Geodiffussr: Generative Terrain Texturing with Elevation Fidelity

ANONYMOUS AUTHOR(S)
SUBMISSION ID: 190

Large-scale terrain generation remains a labor-intensive task in computer graphics. We introduce *Geodiffussr*, a flow-matching pipeline that synthesizes text-guided texture maps while strictly adhering to a supplied Digital Elevation Map (DEM). The core mechanism is multi-scale content aggregation (MCA): DEM features from a pretrained encoder are injected into UNet blocks at multiple resolutions to enforce global-to-local elevation consistency. Compared with a non-MCA baseline, MCA markedly improves visual fidelity and strengthens height–appearance coupling (FID ↓ 49.16%, LPIPS ↓ 32.33%, ΔdCor ↓ to 0.0016). To train and evaluate Geodiffussr, we assemble a *globally distributed, biome- and climate-stratified* corpus of triplets pairing SRTM-derived DEMs with Sentinel-2 imagery and *vision-grounded* natural-language captions that describe visible land cover. We position Geodiffussr as a strong baseline and step toward controllable 2.5D landscape generation for coarse-scale ideation and previz, complementary to physically based terrain and ecosystem simulators. We plan to release code and data to spur research on geometry-conditioned generative pipelines.

CCS Concepts: • **Computing methodologies** → Machine learning approaches; Neural networks; *Machine learning*; **Texturing**; **Modeling and simulation**.

Additional Key Words and Phrases: Terrain texturing, remote sensing, flow matching, multi-scale content aggregation

ACM Reference Format:
Anonymous Author(s). 2025. Geodiffussr: Generative Terrain Texturing with Elevation Fidelity. *ACM Trans. Graph.* 1, 1 (August 2025), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Realistic digital terrains are central to games, virtual production, simulation, and visualization. Practical pipelines must satisfy two competing demands: *geometric fidelity* and *visual richness* aligned with biomes or art direction. Manual authoring (sculpting, mask painting) is labor-intensive; procedural noise helps ideation but offers limited control when strict geometry adherence is required. In this work we explicitly target *coarse-scale terrain ideation*: our experiments operate at 32×32 base textures (≈30 m/px) and include 3D renders for communication.

We introduce *Geodiffussr*, a text-guided, DEM-aware generative pipeline based on flow matching. The central design choice is *multi-scale content aggregation* (MCA): VGG-derived DEM features are injected into UNet blocks at multiple resolutions, providing coarse-to-fine cues about global silhouettes and local ridge/valley structure. We also create a dataset pairing DEMs with satellite imagery and vision-grounded captions of visible land cover (appearance-centric rather than metadata-driven).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM 0730-0301/2025/8-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Positioning and scope. Geodiffussr is intended as a fast, controllable baseline for layout exploration and previz, complementary to physically-based terrain/biome stacks (erosion, sediment transport, snow/dune processes, ecosystem simulators). Our focus is adherence to a supplied DEM under text prompts; scaling to production resolutions is discussed in §5.

In summary, we present the following contributions:

- An open, biome-diverse remote sensing dataset containing triplets of DEMs, satellite images, and natural-language captions.
- A novel flow matching-based generative pipeline leveraging multi-scale content aggregation (MCA) for geometry adhering terrain texturing. Since this deals with a new task done in the text-to-terrain domain, we hope this work will stand as a baseline and support future works.
- Ablations showing a significant performance boost with our proposed method with MCA, as well as scaling with model size.

Fig. 1. **Examples of rendered 2.5D terrains using our proposed approach.** We introduce Geodiffussr, a flow matching-based generative pipeline that can create terrain texture maps from intuitive text prompts, while realistically adhering to a specified Digital Elevation Map (DEM) by leveraging Multi-Scale Content aggregation (MCA). This provides a new baseline for text-conditioned, DEM-aware terrain synthesis and a stepping-stone toward fully controllable landscape generation.

2 RELATED WORK

Text-conditioned diffusion/flow. Diffusion/flow models achieve high-fidelity synthesis with conditioning (text, layout, edges); cross-attention provides semantics and auxiliary encoders inject structure.
Text-to-2.5D terrain. MESA [Borne-Pons et al. 2025] trains a text-to-2.5D pipeline on global DEM–imagery pairs with metadata-driven captions. We instead use *appearance-centric* captions and inject external DEM features via MCA to *enforce* adherence to a provided DEM (we do not generate the DEM).

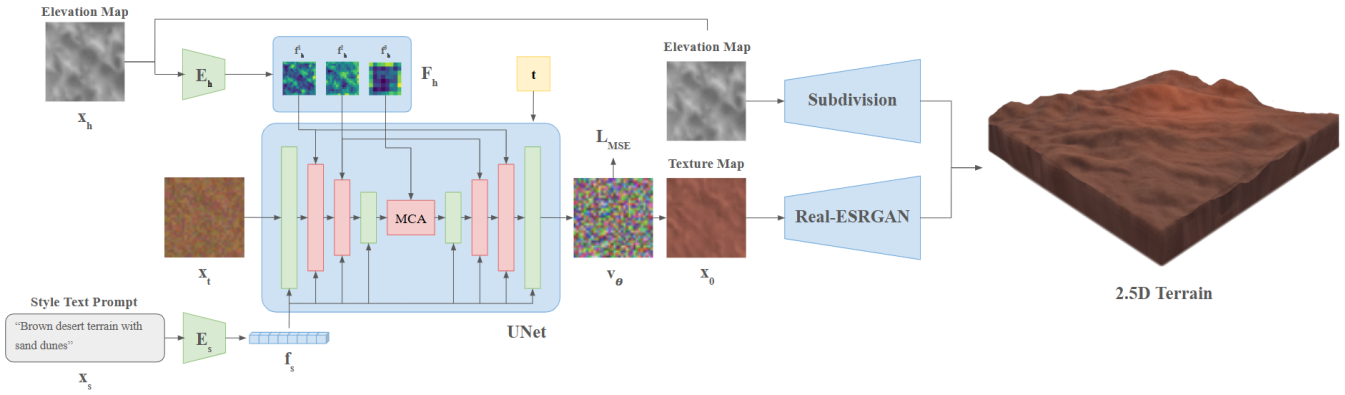

Fig. 2. **Geodiffussr Pipeline.** We condition a flow matching model on both text embeddings and Digital Elevation Maps (DEMs). Specifically for DEMs, we take multi-scale features from a pretrained VGG-16 model and inject into the UNet blocks. The source DEM and generated texture map are increased in resolution via subdivision and Real-ESRGAN superresolution [Wang et al. 2021] respectively for rendering purposes. Combining these results in a 2.5D representation of a terrain as shown on the right.

Physically-based terrain and ecosystems. Hydraulic/thermal erosion, sediment transport, snow/dune/glacier dynamics, and vegetation/ecosystem simulators deliver physical realism and layered materials [Cordonnier et al. 2023; Št'ava et al. 2008; Stomakhin et al. 2013]. Our learned, promptable texturing is complementary: it targets rapid, appearance-centric ideation conditioned on a supplied DEM and can be coupled with physical layers for production assets (§5).

3 METHOD

3.1 Model Architecture

One of the most important aspects of the text-conditioned generation process is that it must adhere to the structure of the user-inputted height map. In order to enforce strict adherence to the height map, we incorporate Multi-scale Content Aggregation (MCA) [Yang et al. 2023] by injecting coarse-to-fine feature information into the UNet of the flow matching model. The motivation behind this conditioning mechanism is that image encoder's intermediate feature representations provide much richer information about the underlying height map than simply using the height map directly. In our implementation, we opted to use a pretrained VGG-16[Simonyan and Zisserman 2015] model and extract several feature maps at the 32x32, 16x16, and 8x8 resolutions and inject those into the Unet using a Squeeze-and-Excitation (SE)[Hu et al. 2018] block. The SE block serves as a channel mixing mechanism to incorporate the information from the VGG feature maps before downsampling it back to the original size.

For the text-conditioning we utilize text embeddings extracted from the final hidden state of the Flan-T5 series (Small, Base, and Large)[et al. 2022], and perform pixel-wise cross attention in a similar fashion to popular diffusion models such as Stable Diffusion. Furthermore, we also incorporate pixel-wise self-attention blocks before performing the conditioning mechanisms.

Figure 2 portrays the overall pipeline architecture detail, where a UNet model is conditioned on both text-embeddings and DEM features. Specifically, the DEM conditioning is done via MCA. After the Geodiffussr model, both DEM and texture are upscaled using simple subdivision and Real-ESRGAN superresolution [Wang et al. 2021] for 3D rendering purposes.

3.2 Dataset

We created a new dataset specialized for text-to-terrain purposes, containing 380K sets of Digital Elevation Maps (DEM), satellite images, and synthetic text labels.

For geospatially diverse data, we first constructed a catalogue of 200+ non-overlapping 1°×1° Areas of Interest (AOIs) that jointly span every major terrestrial biome. We performed biome stratification based on the WWF Terrestrial Ecoregions map and Koppen-Geiger climate classes [Kottek et al. 2006; Olson et al. 2001]. For each of the 16 super-biomes we targeted 10+ representative sites distributed across multiple continents. This was a method we implemented to obtain a smaller representative of the global geographical features without needing for the near to entire area coverage. Based on this catalogue, Digital Elevation Maps are sampled from USGS SRTMGL1 v003 [Farr et al. 2007] and the satellite images are taken from Copernicus Sentinel-2 [Drusch et al. 2012].

Captions are synthesized using a pretrained language model [Leenstra et al. 2021]. Each satellite image is captioned using the Gemini 2.0 Flash Lite model[1], considering its efficiency in multi-modality.

4 EXPERIMENTS

4.1 Ablation Study

We isolate the effect of each design choice by varying one factor at a time while holding the rest fixed (AdamW, lr 5×10^{-4}, Flan-T5-Small for text, CFG scale w=8, etc.)

Firstly, we observed the effect of MCA by comparing three settings: Full MCA (32×32, 16×16, 8×8 VGG features with SE adapters),

[1]Google Gemini 2.0 Flash-Lite is a cost-efficient, low-latency variant of the Gemini 2.0 Flash family. Detailed specs and usage are available in the Vertex AI model garden: https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash-lite

Single MCA (only 16×16 injection), and Non-MCA (direct DEM concatenation). Then, we compared model performance with its size, to see if there is potential for greater performance with increasing model size.

We evaluate Geodiffussr quantitatively and qualitatively, focusing on two axes: (i) texture quality, measured by FID and LPIPS, and (ii) elevation-texture alignment, measured by relative distance correlation (ΔdCor) between the hue/saturation/value channels of the generated texture and the input DEM.

We report four complementary measures:

- **FID ↓** [Heusel et al. 2017]: distributional distance between Inception features of generated vs. real satellite images (perceptual realism).
- **LPIPS ↓** [Zhang et al. 2018]: patch-level perceptual difference to reference imagery.
- **MSE ↓**: per-pixel reconstruction error (L_2).
- **ΔdCor ↓**: absolute gap to the dataset's ground-truth dependence between DEM elevation and HSV(X), i.e., $\mathbf{\Delta dCor} = \left|\mathbf{dCor(HSV}(X)\mathbf{, DEM)} - \mathbf{dCor}_{\mathrm{gt}}\right|$, with $\mathbf{dCor}_{\mathrm{gt}}\mathbf{=0.3816}$ (captures geometry–appearance coupling, including nonlinear dependence).

4.2 Main results

Full MCA achieves FID 10.29, LPIPS 0.066, MSE 0.0166, and ΔdCor 0.0016. Versus a non-MCA baseline, FID drops by 49.16% and LPIPS by 32.33%, while ΔdCor improves from 0.0756 to 0.0016 (closing 97.9% of the remaining gap to the dataset's ground-truth dependence).

4.3 Quantitative Effect of Multi-scale Content Aggregation (MCA)

We compare three geometry injection settings: full MCA, single MCA, and non-MCA baseline (Table 1). When DEM features are fused at three scales (32×32, 16×16, 8×8) via MCA, the model attains its best scores across all metrics (FID 10.29, LPIPS 0.066, MSE 0.0166, and ΔdCor 0.0016) indicating sharper, more geometry-consistent textures. Injecting at only the 16×16 scale yields intermediate performance (FID 14.50, LPIPS 0.085, MSE 0.0144, ΔdCor 0.0196), while removing MCA entirely (direct DEM concatenation) degrades results (FID 20.24, LPIPS 0.0977, MSE 0.0184, ΔdCor 0.0756) severely.

What's especially notable is that full MCA injection closes 97.9 % of the remaining dCor gap to the ground-truth, compared to a non-MCA baseline, demonstrating that multi-scale fusion enforces consistent shading and color transitions reflecting true topography present in the dataset. These findings confirm that multi-scale fusion substantially improves both perceptual fidelity and height–texture correlation.

4.4 Qualitative Effect of Multi-Scale Content Aggregation (MCA)

Figure 3 contrasts outputs produced with and without Multi-scale Content Aggregation (MCA) when both models receive the same DEMs and prompts.

With MCA (center), the texture conforms to the underlying relief at every scale from global to local ones. For instance, with the snowy

Setting	FID ↓	LPIPS ↓	MSE ↓	Δ dCor ↓
Full MCA	**10.29**	**0.066**	0.0166	**0.0016**
Single MCA	14.50	0.085	**0.0144**	0.0196
Non-MCA	20.24	0.098	0.0184	0.0756

Table 1. **Comparison between varying amounts of MCA injections.** Results show that injecting MCA into every dimension improves performance.

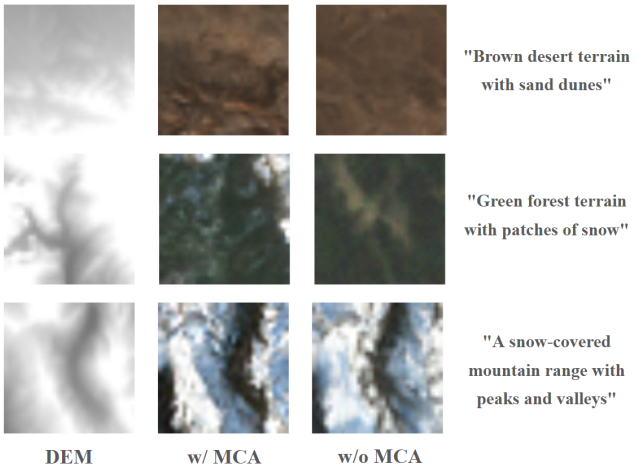
"Brown desert terrain with sand dunes"
"Green forest terrain with patches of snow"
"A snow-covered mountain range with peaks and valleys"
DEM
w/ MCA
w/o MCA

Fig. 3. **Comparison of generated results between Full MCA (center) and non-MCA (right) versions of Geodiffussr.** The textures are generated with various prompts featuring different biomes and a source DEM (left).

mountain range on the bottom row, snow settles cleanly on the ridge tops, darker rock appears at the valley, and subtle gray shadows accentuate minor spur lines. The viewer can infer the approximate height field from the texture alone, confirming that the network has internalized the DEM-to-texture relationship.

Without MCA (right), that correspondence collapses. Again with the same example, a single diagonal band of rock is hallucinated through the center while surrounding areas are indiscriminately coated in snow, ignoring the complex combination of peaks and valleys.

This side-by-side shows that MCA's coarse-to-fine fusion allows the network to internalize both global structural cues (e.g., mountain silhouettes) and fine-scale details (e.g., micro-ridges), whereas removing MCA entirely leads to severe texture collapse.

4.5 UNet Model Size

On the other hand, when comparing the performance of the generative models when varying the UNet model sizes, we observed a consistent increase in the performance of the models as the model capacity grew. These results reveal a promising trend that suggests future work may also benefit from scaling their models to even larger sizes than those we trained on, with little indication of a performance plateau within the 45M, 75M, 102M parameter models

that we have tested. However, in line with established scaling laws, we expect that expanding model size will also require larger training sets to avoid diminishing returns.

Model Size	FID ↓	LPIPS ↓	MSE ↓	Δ dCor ↓
45M	23.08	0.121	0.0235	0.0656
75M	14.50	0.085	**0.0144**	0.0196
102M	**10.29**	**0.066**	0.0166	**0.0016**

Table 2. **UNet Model Size Comparison.** Results show that increasing model size steadily improves performance.

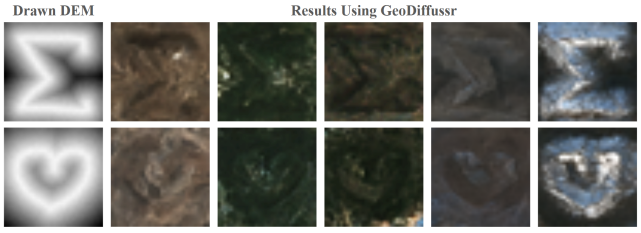

Fig. 4. **Sketch DEMs.** Geodiffussr generalizes to user-drawn synthetic DEMs, producing coherent, prompt-consistent textures. This demonstrates the flexibility of our model to unseen complex geometry, and its potential to be applied with user-guided DEMs.

5 DISCUSSION AND LIMITATIONS

Why MCA works. Injecting DEM features at multiple scales exposes the UNet to both global silhouettes and fine relief, which we find is essential for consistent shading/biome placement relative to elevation.

Scaling to production. Since Geodiffussr is still working with a coarse resolution (32×32), we suggest some approaches to higher resolutions for practical use: (i) a *global-context token* from pooled full-scene DEM features injected into MCA at each scale to keep long-range structure coherent, (ii) a *coarse-to-fine cascade* (32→128→256 px) conditioned on the upsampled prior stage with elevation-aware regularizers (e.g., gradient alignment) to preserve snowlines and drainage, and (iii) a *DEM-aware super-resolution head* tuned for geospatial edges.

Applications and integration. We envision an end-to-end 2.5D pipeline driven by text and sketches: a sketch-to-DEM module converts hand-drawn contours into elevation maps [Hu et al. 2024; Wang and Kurabayashi 2020], and Geodiffussr applies multi-scale, promptable texturing—yielding terrains whose elevation and appearance jointly follow the user's sketch and prompt. A prototype of this idea is illustrated in Figure 4

6 CONCLUSION

Geodiffussr couples text guidance with explicit, multi-scale DEM conditioning for terrain texturing. Our experiments validate MCA's impact on perceptual quality and elevation alignment, establishing a compact, reproducible baseline for controllable 2.5D terrain synthesis. We hope to openly contribute our baseline and dataset to spur future research into realistic, user-guided terrain texturing.

REFERENCES

Paul Borne-Pons, Mikolaj Czerkawski, Rosalie Martin, and Romain Rouffet. 2025. MESA: Text-Driven Terrain Generation Using Latent Diffusion and Global Copernicus Data. arXiv preprint arXiv:2504.07210. https://arxiv.org/abs/2504.07210

Guillaume Cordonnier, Adrien Peytavie, Eric Galin, Eric Guérin, Jérémie Arias, Marie-Paule Cani, and James Gain. 2023. Forming Terrains by Glacial Erosion. *ACM Transactions on Graphics* 42, 6, Article 167 (2023). https://doi.org/10.1145/3618390

Matthias Drusch, Umberto Del Bello, Sophie Carlier, Olivier Colin, Violeta Fernandez, François Gascon, Benoit Hoersch, Christoph Isola, Pietro Laberinti, Philippe Martimort, Alexis Meygret, Fiona O'Connor, Christina Potsiou, Javier Santolaria, Martin Schmidt, Odile Sy, and Paolo Bargellini. 2012. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment* 120 (2012), 25–36. https://doi.org/10.1016/j.rse.2011.11.026

Hyung Won Chung et al. 2022. Scaling Instruction-Finetuned Language Models. *arXiv preprint arXiv:2210.11416* (2022). https://arxiv.org/abs/2210.11416

Tom G. Farr, Paul A. Rosen, Eric Caro, Rodney Crippen, and Riley Duren. 2007. The Shuttle Radar Topography Mission. *Reviews of Geophysics* 45, 2 (2007), RG2004. https://doi.org/10.1029/2005RG000183

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, Vol. 30. https://arxiv.org/abs/1706.08500

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. https://arxiv.org/abs/1709.01507

Zexin Hu, Kun Hu, Clinton Mo, Lei Pan, and Zhiyong Wang. 2024. Terrain Diffusion Network: Climatic-Aware Terrain Generation with Geological Sketch Guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*. https://arxiv.org/abs/2308.16725

M. Kottek, J. Grieser, C. Beck, B. Rudolf, and F. Rubel. 2006. World Map of the Köppen-Geiger Climate Classification Updated. *Meteorologische Zeitschrift* 15, 3 (2006), 259–263. https://doi.org/10.1127/0941-2948/2006/0130

Marrit Leenstra, Diego Marcos, Francesca Bovolo, and Devis Tuia. 2021. Self-supervised Pre-training Enhances Change Detection in Sentinel-2 Imagery. *arXiv preprint arXiv:2101.08122* (2021). https://arxiv.org/abs/2101.08122

D. M. Olson, E. Dinerstein, E. D. Wikramanayake, N. D. Burgess, G. V. N. Powell, E. C. Underwood, J. A. D'Amico, I. Itoua, H. E. Strand, J. C. Morrison, C. J. Loucks, T. F. Allnutt, T. H. Ricketts, Y. Kura, J. F. Lamoreux, W. W. Wettengel, P. Hedao, and K. R. Kassem. 2001. Terrestrial Ecoregions of the World: A New Map of Life on Earth. *BioScience* 51, 11 (2001), 933–938. https://doi.org/10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2

Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations* (2015). https://arxiv.org/abs/1409.1556

Ondřej Šťava, Bedřich Beneš, Michal Brisbin, and Jaroslav Křivánek. 2008. Interactive Terrain Modeling Using Hydraulic Erosion. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '08)*. Eurographics Association, 201–210.

Alexey Stomakhin, Craig Schroeder, Lena Chai, Joseph M. Teran, and Andrew Selle. 2013. A Material Point Method for Snow Simulation. *ACM Transactions on Graphics* 32, 4, Article 102 (2013), 10 pages. https://doi.org/10.1145/2461912.2461948

Tong Wang and Shuichi Kurabayashi. 2020. Sketch2Map: A Game Map Design Support System Allowing Quick Hand Sketch Prototyping. In *IEEE Conference on Games (CoG)*. 596–603. https://doi.org/10.1109/CoG47356.2020.9231924

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. 2021. Real-ESRGAN: Training Real-World Blind Super-Resolution With Pure Synthetic Data. *arXiv preprint arXiv:2107.10833* (2021). https://arxiv.org/abs/2107.10833

Zhenhua Yang, Dezhi Peng, Yuxin Kong, Yuyi Zhang, Cong Yao, and Lianwen Jin. 2023. FontDiffuser: One-Shot Font Generation via Denoising Diffusion with Multi-Scale Content Aggregation and Style Contrastive Learning. arXiv preprint arXiv:2312.12142. https://arxiv.org/abs/2312.12142

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. https://arxiv.org/abs/1801.03924